HHS Public Access

Author manuscript

Field methods. Author manuscript; available in PMC 2019 November 27.

Published in final edited form as:

Field methods. 2019 November ; 31(4): 328–343. doi:10.1177/1525822X19871546.

The Effects of Embedding Closed-ended Cognitive Probes in a Web Survey on Survey Response

Paul J. Scanlon[1]

[1]National Center for Health Statistics, Centers for Disease Control and Prevention, Hyattsville, MD, USA

Abstract

Web, or online, probing has the potential to supplement existing questionnaire design processes by providing structured cognitive data on a wider sample than typical qualitative-only question evaluation methods can achieve. One of the practical impediments to the further integration of web probing is the concern of survey managers about how the probes themselves may affect response to other items and to a questionnaire as a whole. This study explores the effects web probes had on response to a self-administered web survey by comparing two rounds of this survey—one without web probes and one with web probes—that were administered to a probability-based panel of approximately 100,000 American adults. While the item response to the probes themselves appears to be related to the way they are formatted, the findings indicate that web probes do not have an overall negative effect on a questionnaire in which they are embedded.

Introduction

With the maturation of commercially available online panels of survey respondents (typically referred to as “web panels”), online cognitive probing has developed as a new questionnaire evaluation method. Online cognitive probing, or web probing, leverages the larger number of respondents and geographic diversity that web panels can provide to get broader information about question performance than is usually available from small-scale evaluation studies.

Web-probing methodology builds on the use of respondent debriefings and embedding questions or experiments into either field tests or production surveys to evaluate question performance. Early work, such as that by Schuman (1966), Converse and Presser (1986), and Cannell et al. (1989), showed that relatively simple, open-ended probe questions (such

Corresponding Author: Paul J. Scanlon, National Center for Health Statistics, Centers for Disease Control and Prevention, 3311 Toledo Road, Hyattsville, MD 20782, USA. pscanlon@cdc.gov.

Author’s Note

The opinions expressed in this article are my own and do not reflect the view of the Centers for Disease Control and Prevention, the Department of Health and Human Services, or the U.S. Government.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Supplemental Material

Supplemental material for this article is available online.

as “Could you tell me more about that?”) proved to be an efficient way to identify problematic questions within the survey environment. The use of such probes in field tests has always been limited because of the high costs and logistical challenges of adding and processing the items in field settings as well as concerns about how these additional items affect overall survey response. However, embedding cognitive probes in field tests has become more common in the last decade with the advent of Internet mode surveys and readily available panels of respondents (Willis 2014).

While some previous work (i.e., Edgar 2013; Murphy et al. 2014) suggested that web probing could replace traditional face-to-face interviewing in some cases, some recent research has concluded that it should be used to supplement, and not supplant, live cognitive interviews (Behr 2016: Behr et al. 2012; Fowler 2016; Fowler et al. 2015; Russell and Hubley 2016). Typically, these web probes are administered as open-ended questions with text fields, attempting to obtain the same sort of information that face-to-face cognitive interviews can provide (Behr et al. 2017; Meitinger and Behr 2016). These data are usually coded and quantitatively analyzed, but doing so requires substantial effort and introduces the potential for intercoder unreliability and other coding issues.

A more effective way to quantitatively evaluate cognitive question performance is to employ closed-ended probes, which—compared to open-ended probes—are not only less burdensome but also eliminate problems related to intercoder unreliability. They are developed by analyzing qualitative findings, such as those from cognitive interviews, to determine the patterns of response of a particular survey item; the individual patterns within this response schema are then used as the answer categories in the closed-ended probe. Administering these probes on a web survey then allows for the quantitative analysis of patterns across subgroups; if administered on a representative survey, these probes can then be used to determine the distribution of interpretations across a population. For example, following cognitive interviews that found a number of both in- and out-of-scope interpretations for a set of anxiety questions (Miller et al. 2011), a closed-ended probe was designed using these patterns (Miller and Maitland 2010). After being administered in a cross-national field test, Loeb (2016) was able to determine whether certain countries were more likely to use an out-of-scope interpretation than others.

While issues of logistics and costs associated with embedding probes into field tests have been alleviated with the maturation of online survey panels, one impediment that remains is the perception by survey managers that the probes themselves may have negative effects on how respondents interact with the rest of the questionnaire. There are very little empirical data on whether or not cognitive probes cause a framing or reactivity effect (Koskey 2016). While Beatty (2004) presents some evidence from a single cognitive interviewing study that the precision of respondents’ survey answers can be affected by the formatting of probes, no empirical data are currently available on the reactivity of closed-ended probes. It is possible that adding probes to a survey may impact respondents’ approaches and response processes to the other items on the survey; this, in turn, could affect unit and item response as well as the final estimates the survey produces. This study set out to explore whether probe questions have measurable effects on survey response by considering the following three questions:

1. How do closed-ended cognitive web probes affect the overall response and completion rates of surveys in which they are embedded?
2. How do these closed-ended web probes affect responses to nonprobe questions and variables on surveys in which they are embedded?
3. How do respondents answer closed-ended probes themselves?

Study Design

The Web Panel Sample

To investigate the use of closed-ended cognitive web probes for questionnaire design, the National Center for Health Statistics conducted an Internet-administered survey over two rounds, known as the Research and Development Survey (RANDS). The survey was administered to members of the Gallup Panel, a probability-based panel of approximately 100,000 American adults managed by the Gallup Organization, which recruits randomly selected respondents to its daily RDD tracking surveys (Gallup 2016). A stratified sample (based on age, education, and race/ethnicity) was invited to participate in RANDS. The first round was conducted in late 2015 and captured 2,304 complete responses (and an additional 118 partial responses) for a response rate (RR6, American Association for Public Opinion Research [AAPOR] 2016) of 24.69%; the second round was conducted in the beginning of 2016 and captured 2,480 completes (plus an additional 148 partials), for an AAPOR RR6 response rate of 31.93%.

The samples of the two rounds did not overlap, but the samples' demographic compositions were comparable (see Supplemental Table S1). In the file provided by Gallup, "complete" responses are those where respondents made it to the last screen of the survey and clicked the "End Survey" button; if a respondent started the survey but did not click this final link, they were coded as a "partial" response. While the Gallup Panel includes both Internet and non-Internet users, in order to focus specially on web response, RANDS was only administered online, and no special attempts were made to include Gallup Panel members who did not have access to the Internet. However, given that it is based on a statistical sample in the first place, Gallup was able to construct poststratification weights, which took age, race, ethnicity, sex, and educational attainment into account. Due to this ability to weight the sample to the population, we expect results from a recruited panel such as the Gallup Panel to provide more reliable information about potential American survey respondents than an opt-in survey would.

The Survey Questions

The first round's questionnaire consisted of 72 National Health Interview Survey (NHIS) questions, while the second round included both these NHIS questions as well as 21 closed-ended web probes. Topics included general health, food security, work status, chronic conditions, health behaviors, physical activity, health-care access and use, psychological distress, Internet use, and anxiety (see Scanlon [2017] for full questionnaires).

The process used to develop these probes is discussed in detail elsewhere (see Scanlon 2016, forthcoming); in short, 30 cognitive interviews in the Washington, DC, metropolitan area

were analyzed in an effort to find the full set of patterns of interpretation that respondents used while answering the NHIS questions that were included on the first round's questionnaire. These patterns then served as the basis for the answer categories in the closed-ended web probes. For example, one of the RANDS items was the "effort" question in the Kessler 6 Psychological Distress Scale (Kessler et al. 2003), which reads:

> In the past 30 days, how often did you feel like everything was an effort? All, Most, Some, a Little, or None of the Time?

Cognitive interviewing revealed that some respondents interpreted the idea that "everything was an effort" to be a positive, motivating force, and others believed it to be something negative, while a few believe that it was a neutral feeling. Given this set of patterns of interpretation, a probe was designed for the RANDS questionnaire following this effort question that read:

> Would you consider everything being an effort to be a good thing, a bad thing, or neither good nor bad?

The final questionnaire, including the probe questions, was tested ($n = 5$ cognitive interviews) to confirm that each probe captured the aspect of the question response process for which it was designed.

Results

Results are presented here in order of the research questions stated above. Unless otherwise noted, the data presented throughout this article are weighted, and analyses exclude nonrespondents and partial respondents. All statistical analyses were conducted using R software Version 3.5.1. A significance level of 0.05 is used throughout.

Both rounds 1 and 2 were sampled from the same panel population, but only round 2 included probes; this two-round structure provides an opportunity to examine how the presence of web probes affects response behavior. Round 2 was administered four months after round 1 to leverage the quarterly release of NHIS results; therefore, effects due to this difference in time (such as potential seasonality effects surrounding concepts like health insurance coverage) cannot be disentangled from the effects of the probes.

Question 1: Effects of Probes on Response and Completion Rates

Three metrics were analyzed to explore the impact of the probes on the overall response and completion of the survey: unit completion, breakoff, and overall item response rates.

Unit completion rates.—Gallup's disposition coding scheme does not truly relate to response behavior, since, for example, a respondent could refuse to answer all the questions in the survey but click on this final link and be counted as a "complete." To derive an empirically based disposition code classification scheme following the suggested criteria outlined by AAPOR (2016:14), respondents who provided answers to 80% or more of the questions to which they were eligible were coded as complete cases, while those who answered less than 80% were coded as "partial" cases. There is no significant difference between the unweighted percent of complete (vs. partial) responses in the two rounds of data

collection, with $\chi^2(1, N = 5{,}050) = 1.317$, $p = .251$ (see Supplemental Figure S1a; please note that unweighted data shown as Gallup only assign weights to complete cases).

Breakoffs.—While there are no standard definitions of a survey response breakoff (AAPOR 2016), in this analysis, breakoff status is based on whether respondents answered the final two questionnaire pages administered to them. (The number of questions that are included on these final two pages ranged from two to three questions, based on their individual response path and skip patterns.) No significant difference is observed between the breakoff rates of those who were administered probes and those who were not, with $\chi^2(1, N = 4{,}784) = 3.313$, $p = .069$ (see Supplemental Figure S1b).

Overall item response.—The mean percent item response, across all the survey questions in round 1 is 97.983% ($SD = 3.871$), while the mean in round 2 is 97.862% ($SD = 2.408$). A two-sample t-test indicates that the overall item response between round 1 and round 2 does not differ significantly ($t[164] = 1.245$, $p = .215$).

Question 2: Effects of Probes on Nonprobe Responses

In principal, probes can impact responses to other questions in two ways: through "framing," which could alter the estimate that questions provide when preceded by probes, and by increasing the item missing rate of non-probe questions because of the increased burden they may levy on respondents.

Difference in estimates.—Figure 1 plots the effect sizes of the variables shared by both rounds of RANDS (ordered on the x-axis simply in terms of their position on the questionnaire) between the two rounds using Cohen's d (for continuous and ordinal variables) and Cohen's h (for binary variables). All nominal variables were dichotomized for each answer category and analyzed as binary variables using Cohen's h. For both of these statistics, no difference between the estimates from round 1 and round 2 would produce a statistic that equals zero. Cohen (1988:12–13, 25) notes that whether or not an effect size is meaningful should be left to subject matter experts; however, he provides rough equivalences between effect size and qualitative magnitude. Because the variables under analysis here range over a wide number of subject areas, these rough guidelines are used in this analysis. All the estimates of differences fall within the band of effect sizes that Cohen described as generally negligible for both population means and proportions ($|d| < 0.2$ and $|h| < 0.2$), which is depicted as the shaded area in Figure 1.

Position of nonprobe questions in questionnaire.—To investigate the impact the added burden probes have on the response rates of other questions, the item response rates of questions based on their position in relation to probes can be examined. A broad examination of this position effect can compare questions that follow probes either on the same or next page as compared to questions that do not follow probes. Those questions that do follow probes display a lower percent of item response ($M = 97.011\%$, $SD = 1.121$) as compared to questions that do not follow probes ($M = 98.576\%$, $SD = 9.875$). This difference in means is significant but small ($MD = 1.57$, $t[2{,}479] = 11.51$, $p < .001$, $d = 0.23$). This pattern holds true when this broad measure is broken down further by comparing

questions administered immediately after a probe versus those that do not directly follow a probe ($MD = 1.70$, $t[2,479] = 9.43$, $p < .001$, $d = 0.19$) as well as comparing questions that appear on the screen following a probe versus those that do not ($MD = 1.99$, $t[2,479] = 13.08$, $p < .001$, $d = 0.26$).

Question 3: Response to Cognitive Probes

Respondents in round 2 answered the probes at a lower rate ($M = 87.972\%$, $SD = 10.638$) than they did the nonprobe questions ($M = 97.889\%$, $SD = 11.812$). This difference is significant with a moderate effect ($t[2,479] = -49.512$, $p < .001$, $d = 0.488$), indicating that the probe questions do differ from the nonprobes in some way. Two factors that may contribute to this difference are topic fatigue and the format of the probe questions.

Topic fatigue.—Topic fatigue is the potential that as some respondents answer more questions on the same topic, the more likely they are to refuse future questions about that topic. As shown in Supplemental Table S2, however, only a weak (defined following Cohen [1988, 1992] as an $r < |0.3|$) negative correlation ($r[93]$ −0.240, $p = .019$) exists between the round 2 items' response rates and how long their individual topic sections were. This weak correlation between item response and section length persists when looking at both the probe questions alone ($r[19] = -0.142$, p .539) and the nonprobe items alone ($r = [74] = -0.291$, $p = .012$).

Question format.—As compared to the weak correlations presented above between response and section length, stronger correlations exist when considering the relationship between probe response and the format of the probes, including the number of answer categories ($r[19] = -.757$, $p < .001$) and how the answer categories were structured—as either open-ended, forced-choice, or select-all-that-apply questions ($r[19] = 0.573$, $p = .007$). The number of answer categories used on RANDS' probes ranged from a minimum of three to a maximum of 10, with a rounded mean of 5. (For comparison purposes, the number of answer categories on the nonprobe items ranged from 0 for open-ended questions to a maximum of 10, with a rounded mean of 3.) Of the 21 probe questions, 9 were formatted as forced-choice questions (where respondents could only answer a single answer category), while 12 were formatted as select-all-that-apply questions. The most common format of the web probes was a select-all-that-apply question with four or more answer categories, while the most common format for a nonprobe question was a forced-choice question with two answer categories.

Supplemental Table S2 displays the correlations between the item response rates to all round 2 items together and round 2 probe and nonprobe items with section length, the number of answer categories, and the answer category formats. Descriptively, the difference between the mean percent response to forced-choice ($M = 98.791\%$, $SD = 11.579$) and select-all-that-apply formatted probes ($M = 83.699\%$, $SD = 31.020$) is significant and large ($t[2,479] = 25.101$, $p < .001$, $d = 0.504$). However, the mean rate of response between forced-choice probes and forced-choice nonprobes ($M = 98.553$, $SD = 9.498$) does not differ significantly ($t[2,479] = 1.614$, $p = .107$, $d = 0.032$). Only one nonprobe question was formatted as a

select-all question, and no probes in RANDS used an open-ended format, so those comparisons cannot be made.

Discussion

This article examines three research questions surrounding the use of closed-ended cognitive probes in web surveys: How do they affect overall response? How do they affect response to the nonprobe questions with which they share the questionnaire? and How do respondents interact with the probes themselves? Overall, it appears as though web probes do not negatively affect respondents' interactions with other survey questions, indicating that they are not only able to provide usable information about survey participants' response processes but can also do so in a way that maintains the integrity of the survey in which they are embedded.

Probes' Effects on Response Rates

The most basic question to ask when considering the impact of the probe questions on the rest of the survey instrument is whether their presence made it more or less likely that respondents would complete the survey. The design of RANDS permits an examination of this effect, since the round 1 questionnaire only included NHIS items, while the second round's instrument included both NHIS items and web probes.

If probes cause a significant amount of response burden, the expectation would be that both the overall unit and item response rates in the questionnaire with probes would be lower than in the questionnaire without probes. The analysis of the survey's disposition codes, however, indicated no significant difference in unit response between the rounds without and with probes. Similarly, the questionnaire breakoff rates—which focus on respondents' behaviors across a much more specific set of actions as compared to the disposition codes and are therefore a potentially more precise metric to explore the effects of the web probes on response behaviors—did not differ significantly between the round with and without the probes. These findings indicate that the extra burden added by the probes did not cause respondents to "give up" and stop answering the questionnaire along the way.

While examining the complete/partial and breakoff rates gives a sense of probes' impacts on response behavior, particularly in the final parts of a survey questionnaire, looking at item response rates across the entire questionnaire provides a more holistic view. As before, comparing the item response rates of the questions in round 1 to those in round 2 should indicate whether the presence of the probes decreases survey participation. If this is true, one would expect a higher average item response in first round as compared to the second one. However, this is not the case. Overall, then, considering all three of these data points—survey disposition, breakoff behavior, and item response—it appears as though the presence of web probes does not adversely affect whether respondents answer the items on a questionnaire or complete the survey.

Probes' Effects on Response to Nonprobe Items

Although it does not appear that web probes affect the overall amount of response to a survey, another important methodological consideration to their use is whether, and how, the

probes themselves affect respondents' approaches to the rest of the questions on the survey. As noted above, this impact could theoretically come in two forms: either by framing the respondents' understanding of the questions on the survey (as compared to their understanding of the same questions were the probes not present) or by impacting the item response rates of individual questions. The two-round structure of RANDS permits the examination of the first of these potential effects, while exploring the effects of question position within the second round allows analysis of the latter.

Relying again on the fact that round 1 of the RANDS was administered without any probe questions, while round 2 included them, it is possible to see whether the presence of probes had a meaningful framing impact by examining the estimates the survey data produce. Seventy-eight variables were shared by both the round 1 and round 2 questionnaires. The effect sizes for all of these variables fall well within the range of −0.2 to 0.2, indicating the differences in the estimates between the rounds are negligible, using the rough criteria outlined by Cohen (1988). It would be ideal to use a separate determination on the magnitude of the effect size for each variable based on the advice of each subject's literature and experts. However, this imprecise limit of $|x| < 0.2$ is used here to look across the entire set of variables, which represent over a dozen subject areas. Nonetheless, given both this imprecision and the analytic limitations related to the fact that RANDS was not set up as a true experiment (and thus there may be timing and panel composition effects between the rounds that could affect the estimates), these findings suggest that the presence or absence of web probes does not frame the rest of the items on the survey enough to impact estimates.

Besides the concern about the framing effects web probes may have on survey estimates, another worry is that they increase the burden on respondents (and correspondingly cause them to break off or refuse to answer future items). As noted above, no evidence exists that the closed-ended probes effect breakoff rates. However, there is some evidence that probes may affect the item response and nonresponse rates of the questions they precede. The average item response rates of questions that follow probes are significantly lower than that for questions that do not follow probes. These significant differences persist when comparing both the subsets of questions that immediately follow (vs. those that do not immediately follow) a probe and questions on the page following (vs. those not on a page immediately following) a probe. However, the effect sizes for these three sets of questions are all small (Cohen's $d < 0.3$). The burden of the probe questions (most of which were formatted as "select-all-that-apply" questions with more than two answer categories) therefore may have had a small, latent effect on a respondent's willingness to answer the next few questions.

Given the data available from RANDS, however, it is impossible to disentangle whether this effect is due to the fact that the probes asked about cognitively complex topics (i.e., the question response process) or whether it is due to the greater burden the specific formatting of the web probes placed on the respondents. More specifically designed research is necessary on this point.

Respondents' Interactions with the Web Probes

On their face, closed-ended web probes should appear and function the same as nonprobe items: They are not labeled differently; they ask about similar topics; and they use closed-ended answer categories to elicit respondents' answers. The expectation, therefore, would be that respondents answered the probes at the same rate as the nonprobes. However, the round 2 RANDS data do not support this expectation—probes were answered at a significantly lower rate than the NHIS questions. Two potential factors that may contribute to the observed difference in item response rate between probes and nonprobes are topic fatigue and the formatting of the probes themselves.

As noted above, topic fatigue is the idea that as a respondent receives more questions about a topic, they will become more likely to refuse to answer similar questions. Again, it was very rare across both rounds of data collection for a respondent to simply stop answering the survey questionnaire and quit the survey. Rather, respondents who refused or missed a question tended to continue answering the questionnaire on the next screen—a behavior that suggests topic fatigue. Because the web probes, by design, are located at or near the end of a questionnaire section, their response rates are more likely to be affected by topic fatigue than nonprobe questions. If this is the case, the expectation is that by adding probes to a section of the questionnaire, the fatigue will be exacerbated and would result in increased amount of item nonresponse. However, the RANDS data do not bear this hypothesis out, with only a weak correlation found between section length and probe response rate. Furthermore, there does not appear to be a strong link between the presence of a probe in a questionnaire section and the response to the nonprobe questions in that section. These data indicate that topic fatigue does not contribute meaningfully to the differences observed between the response to probes and nonprobes.

Stronger relationships exist between item response and probe format, indicating that the difference in item response between the probe and nonprobe questions may be largely related to how the web probes are presented (and, correspondingly, to the fact that the probes on the RANDS tended to be formatted differently than the nonprobes). The RANDS data indicate that the average item response rate to probes formatted as select-all questions is lower than that to forced-choice probes and that as the number of answer categories increased, the item response rate dropped. However, it appears as though this effect is largely related to format, and not to the fact that the questions are probes, as no significant difference in item response was observed between forced-choice probes and nonprobes.

The data on both topic fatigue and the effects of formatting on response rates suggest that the probes' formatting is the main culprit in the differential response rate between them and the other survey items and not some inherent cognitive property of probes themselves. As noted previously, the probes used in this project were developed directly from previous qualitative findings and were themselves cognitively evaluated, which likely reduced their burden and related item nonresponse as compared to probes that were not designed through this rigorous process. However, the survey on which this analysis is based did not incorporate controlled experiments that could disentangle design aspects such as question length, the number and format of answer categories, placement within the questionnaire, and topic length. Additionally, all the probes used in this analysis focused on the comprehension

stage of the question response process, and other areas of response such as recall and judgment were not probed; it is possible that probes asking about these areas of response would behave differently than “comprehension probes.” Future methodological research should attempt to design and embed experiments into web surveys to explore these various aspects of design to advance the reliability of web probes.

Web probes, and specifically closed-ended web probes, provide question evaluators with an opportunity to move the analysis of patterns of interpretation from small-scale samples to large and potentially nationally representative ones. By combining quantitative analysis from web probes, particularly of subgroups that may be difficult to recruit for face-to-face interviews, with qualitative analysis from cognitive interviews, survey programs can target their question and questionnaire evaluation more effectively—leading to better survey outcomes. However, one of the largest concerns survey managers have about embedding extra elements into a survey such as web probes is that their presence will negatively affect survey response by increasing burden, item, or unit nonresponse or by framing respondents’ answers to the survey’s core items. Before web probes can be adopted widely as part of the normal survey evaluation processes or even embedded in production surveys, this concern must be assuaged. The analysis of RANDS data presented here indicate that, for closed-ended probes at the very least, the addition of embedded cognitive probes does not negatively affect respondents’ experiences or survey outcomes; given this, survey managers and evaluators should consider closed-ended web probing a viable method for expanding their questionnaire evaluation efforts going forward.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The author would like to acknowledge both the staff at NCHS and participants in the Online Probing sessions at the Seventh Conference of the European Survey Research Association conference in Lisbon who saw earlier drafts and provided valuable feedback.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

References

American Association for Public Opinion Research. 2016 Standard definitions: Final dispositions of case codes and outcome rates for surveys, 9th ed. Chicago: American Association for Public Opinion Research.

Beatty P 2004 The dynamics of cognitive interviewing In Methods for testing and evaluating survey questionnaires, eds. Presser S, Rothgeb JM, Couper MP, Lesser JT, Martin E, Martin J, and Singer E, 45–66. Hoboken, NJ: John Wiley & Sons.

Behr D. Cross-cultural web probing and how it can enhance equivalence in cross-cultural studies. Paper presented at the International Conference on Questionnaire Design, Development, Evaluation, and Testing (QDET2); Miami, FL. November 9–13; 2016.

Behr D, Kaczmirek L, Bandilla W, and Braun M. 2012 Asking probing questions in web surveys: Which factors have an impact on the quality of responses? Social Science Computer Review 30:487–98.

Behr D, Meitinger K, Braun M, and Kaczmirek L. 2017Web probing—Implementing probing techniques from cognitive interviewing in web surveys with the goal to assess the validity of survey questions. Mannheim, Germany: GESIS–Leibniz Institute for the Social Sciences.

Cannell C, Fowler F, Kalton G, Oksenberg L, and Bischoping K. 1989 New quantitative techniques for pretesting surveys. Paper presented at the 47th International Statistical Institute, Paris, France, August 29–September 6.

Cohen J 1988 Statistical power analysis for the behavioral sciences, 2nd ed. Hillsdale, NJ: Lawrence Erlbaum Associates.

Cohen J 1992 A power primer. Psychological Bulletin 112:155–59. [PubMed: 19565683]

Converse JM, and Presser S. 1986 Survey questions: Handcrafting the standardized questionnaire. Newbury Park, CA: SAGE.

Edgar J. Self-administered cognitive interviewing. Paper presented at the 68th AAPOR Conference; Boston, MA. May 16–19; 2013.

Fowler S. The practice of cognitive interviewing through web probing. Paper presented at the International Conference on Questionnaire Design, Development, Evaluation, and Testing (QDET2); Miami, FL. November 9–13; 2016.

Fowler S, Willis G, Moser R, Ferrer R, and Berrigan D. 2015 Use of Amazon MTurk online marketplace for questionnaire testing and experimental analysis of survey features. Paper presented at the 2015 Federal Committee on Statistical Methodology Research Conference, Washington, DC, December 1–3.

Gallup. 2016 Gallup panel whitepaper brief Washington, DC: The Gallup Organization GallupPanel@Gallup.com (accessed March 8, 2017).

Kessler R, Barker PR, Colpe LJ, Epstein JF, Gfroerer JC, Hiripi E, Howes MJ, et al. 2003 Screening for serious mental illness in the general population. Archives of General Psychiatry 60:184–89. [PubMed: 12578436]

Koskey KLK 2016 Using the cognitive pretesting method to gain insight into participants' experiences: An illustration and methodological reflection. International Journal of Qualitative Methods 15:1–13.

Loeb M 2016 Development of disability measures for surveys: The Washington group extended set on functioning In International measurement of disability purpose, method and application, ed. Altman B, 97–122. Cham, Switzerland: Springer International.

Meitinger K, and Behr D. 2016 Comparing cognitive interviewing and online probing: Do they find similar results? Field Methods 28:363–80.

Miller K, and Maitland A. 2010 A mixed-method approach for measurement construction for cross-national studies. Paper presented at the 2010 Joint Statistical Meetings, Vancouver, Canada, July 31–August 5.

Miller K, Mont D, Maitland A, Altman B, and Madans J. 2011 Results of a cross-national structured cognitive interviewing protocol to test measures of disability. Quality and Quantity 45:801–15.

Murphy J, Edgar J, and Keating M. 2014 Crowdsourcing in the cognitive interviewing process. Paper presented at the 69th AAPOR Conference, Anaheim, CA, May 15–18.

Russell L, and Hubley A. 2016 Some thoughts on gathering response process validity evidence in the context of online measurement and the digital revolution In Understanding and investigating response processes in validation research, eds. Zumbo B and Hubley A, 229–50. Cham, Switzerland: Springer International.

Scanlon P. Using web panels to quantify the qualitative. Paper presented at the 71st AAPOR Conference; Austin, TX. May 12–15; 2016.

Scanlon P 2017 Evaluation of the 2015–2016 research and development survey. Hyattsville, MD: National Center for Health Statistics https://wwwn.cdc.gov/qbank/report/Scanlon_2017_NCHS_RANDS.pdf (accessed November 11, 2018).

Scanlon P Forthcoming. Using targeted embedded probes to quantify cognitive interviewing findings In Advances in questionnaire design, development, evaluation and testing, eds. Beatty P, Collins D, Kaye L, Padilla J, Willis G, and Wilmot A. Hoboken, NJ: Wiley (Expected January 2020).

Schuman H 1966 The random probe: A technique for evaluating the validity of closed questions. American Sociological Review 31:218–22. [PubMed: 5910069]

Willis G 2014 Analysis of the cognitive interview in questionnaire design. New York: Oxford University Press.

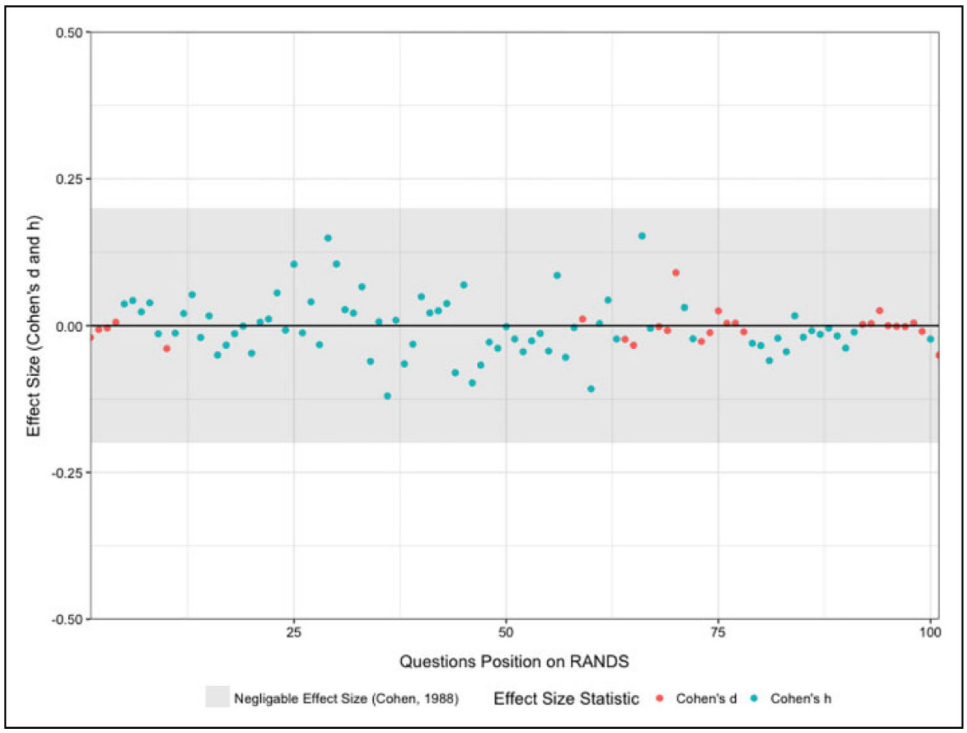

Figure 1.
Effect sizes of the differences in estimates for variables shared across rounds 1 and 2 by item position in the questionnaire.